Performance of Future High-End Computers

David H. Bailey

NERSC Chief Technologist

Lawrence Berkeley National Laboratory

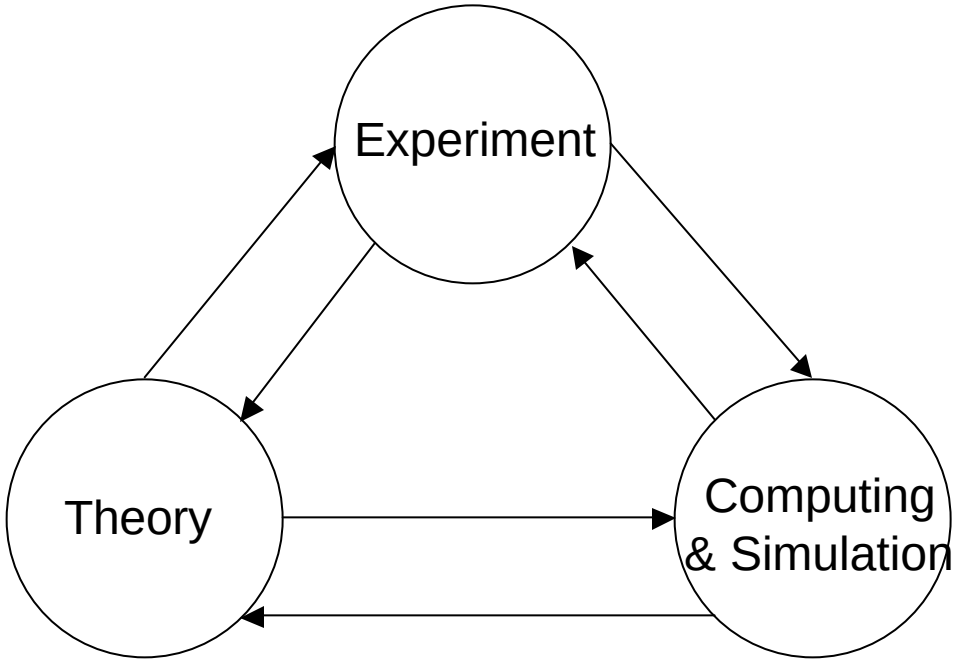
http://www.nersc.gov/~dhbailey

Laplace Anticipates Modern High-End Computers

"An intelligence knowing all the forces acting in nature at a given instant, as well as the momentary positions of all things in the universe, would be able to comprehend in one single formula the motions of the largest bodies as well as of the lightest atoms in the world, provided that its intellect were sufficiently powerful to subject all data to analysis; to it nothing would be uncertain, the future as well as the past would be present to its eyes."

-- Pierre Simon Laplace, 1773

Computing: the Third Mode of Scientific Discovery



Numerical simulations: experimentation by computation.

Who Needs High-End Computers?

Expert predictions:

- (c. 1945) Thomas J. Watson (CEO of IBM):
  "World market for maybe five computers."
- (c. 1975) Seymour Cray:
  "Only about 100 potential customers for Cray-1."
- *(c. 1977) Ken Olson (CEO of DEC):*
  "No reason for anyone to have a computer at home."
- (c. 1980) IBM study:
  "Only about 50 Cray-1 class computers will be sold per year."

Present reality:

- Many homes now have 5 Cray-1 class computers.
- Latest PCs outperform 1988-era Cray-2.

Evolution of High-End Computing Technology

1950	Univac-1	1 Kflop/s (10^3 flop/sec)
1965	IBM 7090	100 Kflop/s (10^5 flop/sec)
1970	CDC 7600	10 Mflop/s (10^7 flop/sec)
1976	Cray-1	100 Mflop/s (10^8 flop/sec)
1982	Cray X-MP	1 Gflop/s (10^9 flop/sec)
1990	TMC CM-2	10 Gflop/s (10^{10} flop/sec)
1995	Cray T3E	100 Gflop/s (10^{11} flop/sec)
2000	IBM SP	1 Tflop/s (10^{12} flop/sec)
2002	Earth Simulator	40 Tflop/s (4×10^{12} flop/sec)

Life Cycle of Scientific Applications

- Infeasible – much too expensive to consider.
- First sketch of possible computation.
- First demo on state-of-the-art highly parallel system.
- Code is adapted for production large-scale runs.
- Code runs on a shared memory multiprocessor.
- Code runs on a single-CPU workstation.
- Code runs on personal computer system.
- Code is embedded in web-based facility.
- Code is embedded in hand-held application.

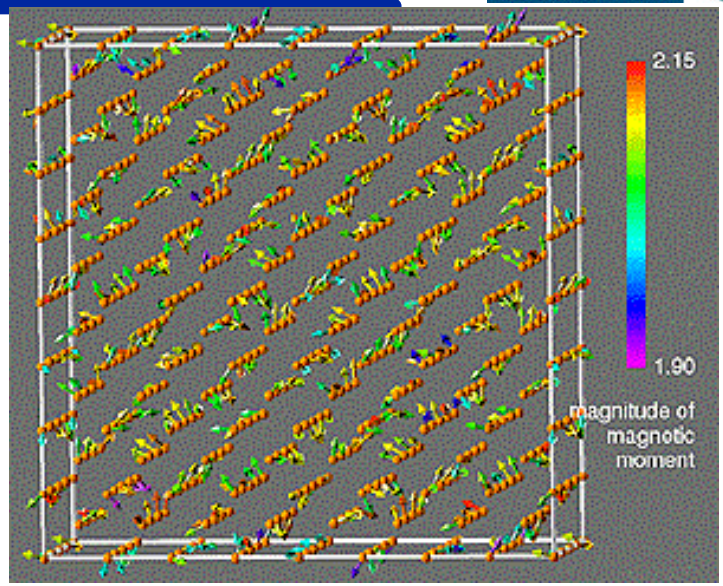
NERSC-3 (Seaborg) System

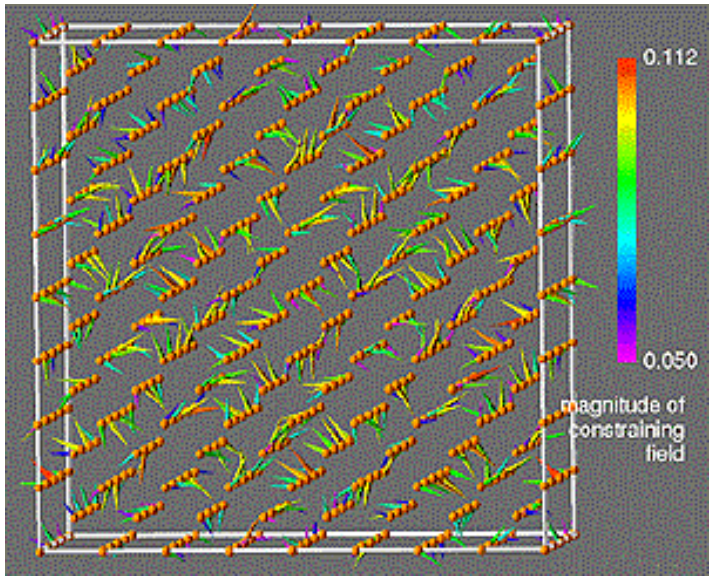
- 6000-CPU IBM SP: 10 Tflop/s (10 trillion flops/sec).
- Currently the world's 4th most powerful computer.

DOE Applications: Materials Science

- 1024-atom first-principles simulation of metallic magnetism in iron was 1998 Gordon Bell Prize winner -- first real scientific simulation to top 1Tflop/s.

- 2016-atom simulation now runs on the NERSC-3 system at 2.46 Tflop/s.





Materials Science Requirements

Electronic structures:

- Current: ~300 atom: 0.5 Tflop/s, 100 Gbyte memory.
- Future: ~3000 atom: 50 Tflop/s, 2 Tbyte memory.

Magnetic materials:

- Current: ~2000 atom: 2.64 Tflop/s, 512 Gbytes memory.
- Future: hard drive simulation: 30 Tflop/s, 2 Tbyte memory.

Molecular dynamics:

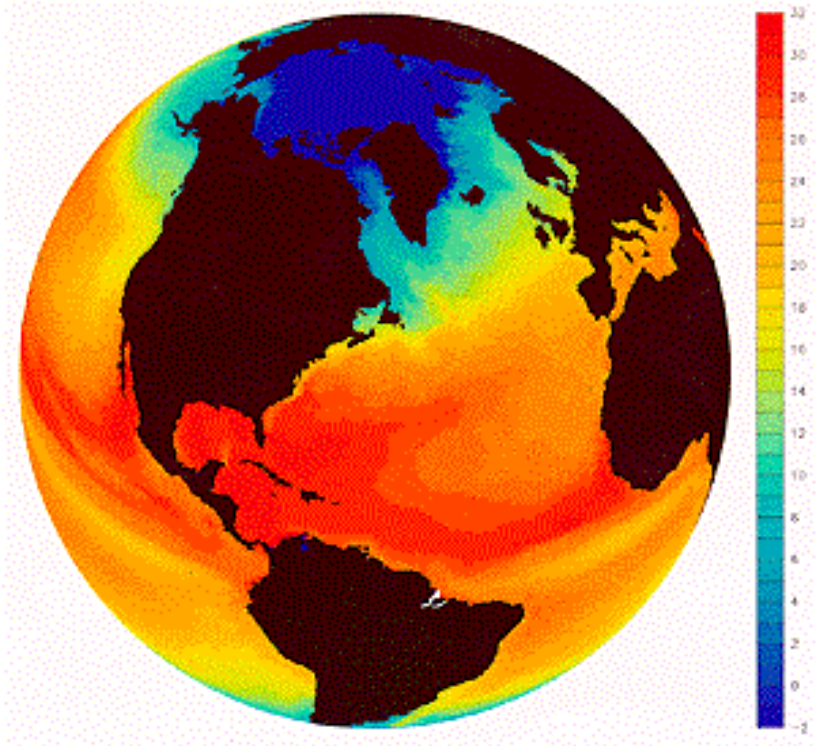
- Current: 10^9 atoms, ns time scale: 1 Tflop/s, 50 Gbyte mem.
- Future: alloys, us time scale: 20 Tflop/s, 4 Tbyte memory.

Continuum solutions:

- Current: single-scale simulation: 30 million finite elements.
- Future: multiscale simulations: 10 x current requirements.

DOE Applications: Environmental Science

Parallel climate model (PCM) simulates long-term global warming.



Climate Modeling Requirements

Current state-of-the-art:

- Atmosphere: 1 x 1.25 deg spacing, with 29 vertical layers.
- Ocean: 0.25 x 0.25 degree spacing, 60 vertical layers.
- Currently requires 52 seconds CPU time per simulated day.

Future requirements (to resolve ocean mesoscale eddies):

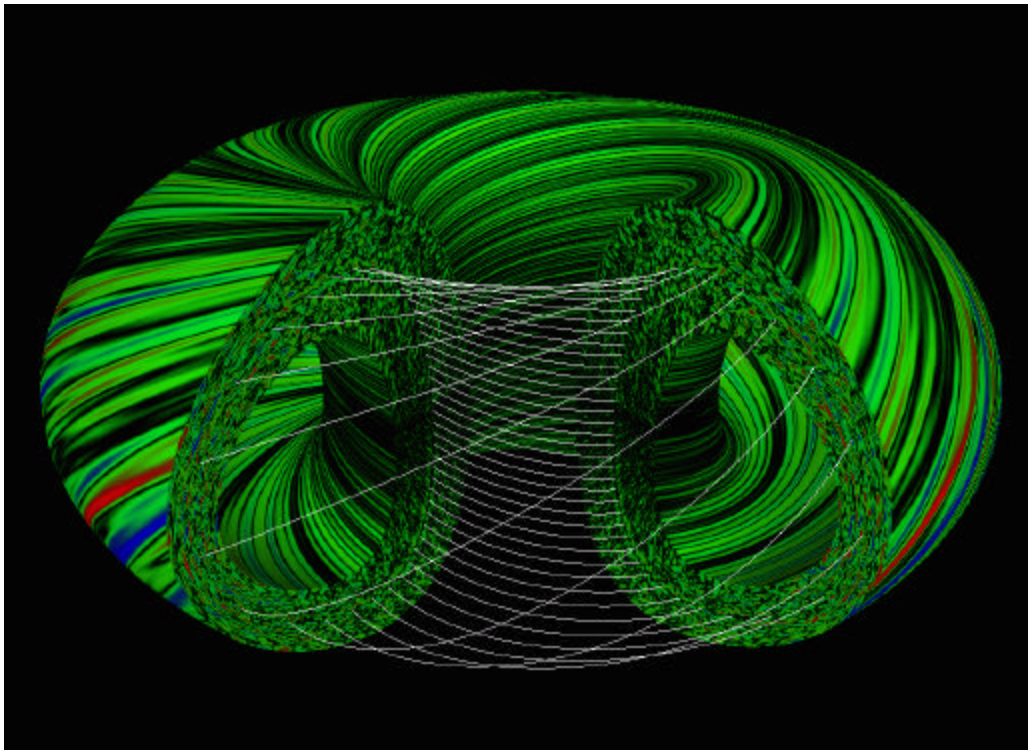
- Atmosphere: 0.5 x 0.5 deg spacing.
- Ocean: 0.125 x 0.125 deg spacing.
- Computational requirement: 17 Tflop/s.

Future goal: resolve tropical cumulus clouds:

- 2 to 3 orders of magnitude more than above.

DOE Applications: Fusion Energy

Computational simulations help scientists understand turbulent plasmas in nuclear fusion reactor designs.



Fusion Requriements

Tokamak simulation -- ion temperature gradient turbulence in ignition experiment:

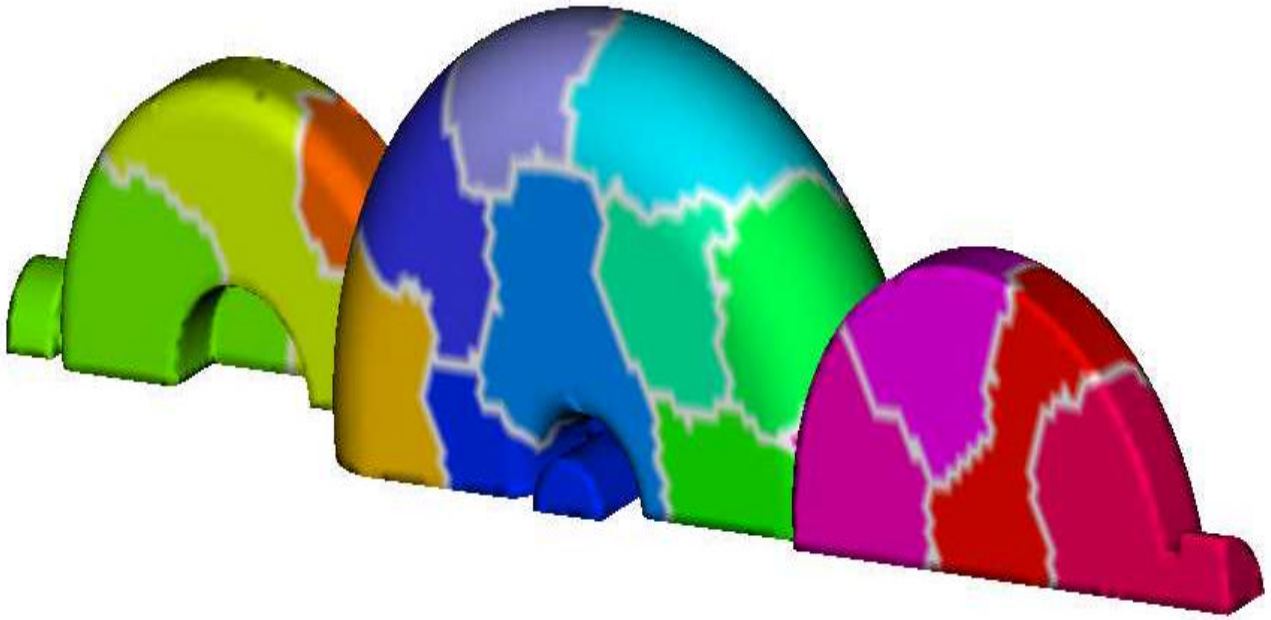
- Grid size: 3000 x 1000 x 64, or about 2 x 10^8 gridpoints.
- Each grid cell contains 8 particles, for total of 1.6 x 10^9.
- 50,000 time steps required.
- Total cost: 3.2 x 10^{17} flop/s, 1.6 Tbyte.

All-Orders Spectral Algorithm (AORSA) – to address effects of RF electromagnetic waves in plasmas.

- 120,000 x120,000 complex linear system.
- 230 Gbyte memory.
- 1.3 hours on 1 Tflop/s.
- 300,000 x 300,000 linear system requires 8 hours.
- Future: 6,000,000 x 6,000,000 system (576 Tbyte memory), 160 hours on 1 Pflop/s system.

NERSC/DOE Applications: Accelerator Physics

Simulations are being used to design future high-energy physics research facilities.



Accelerator Modeling Requirements

Current computations:

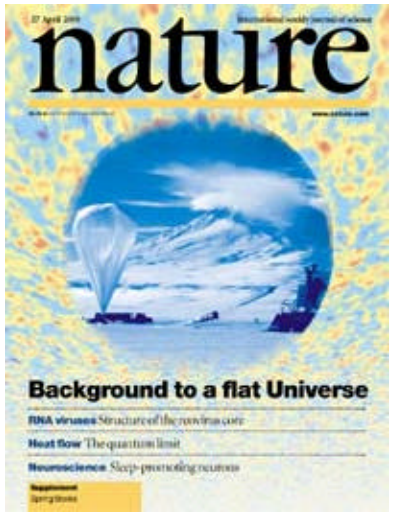
- 1283 to 5123 cells, or 40 million to 2 billion particles.
- Currently requires 10 hours on 256 CPUs.

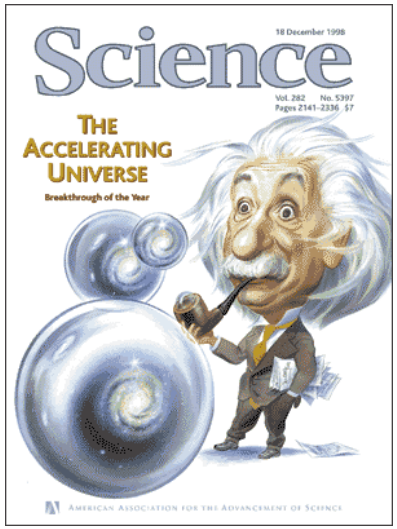
Future computations:

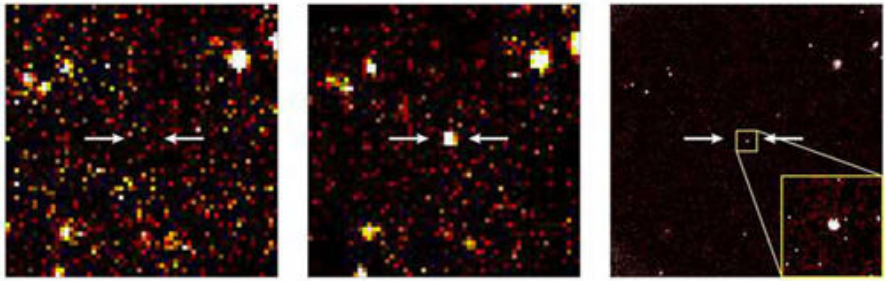
- Modeling intense beams in rings will be 100 to 1000 times more challenging.

DOE Applications: Astrophysics and Cosmology

- The oldest, most distant Type 1a supernova confirmed by computer analysis at NERSC.
- Supernova results point to an accelerating universe.
- Analysis at NERSC of cosmic microwave background data shapes concludes that geometry of the universe is flat.







Astrophysics Requirements

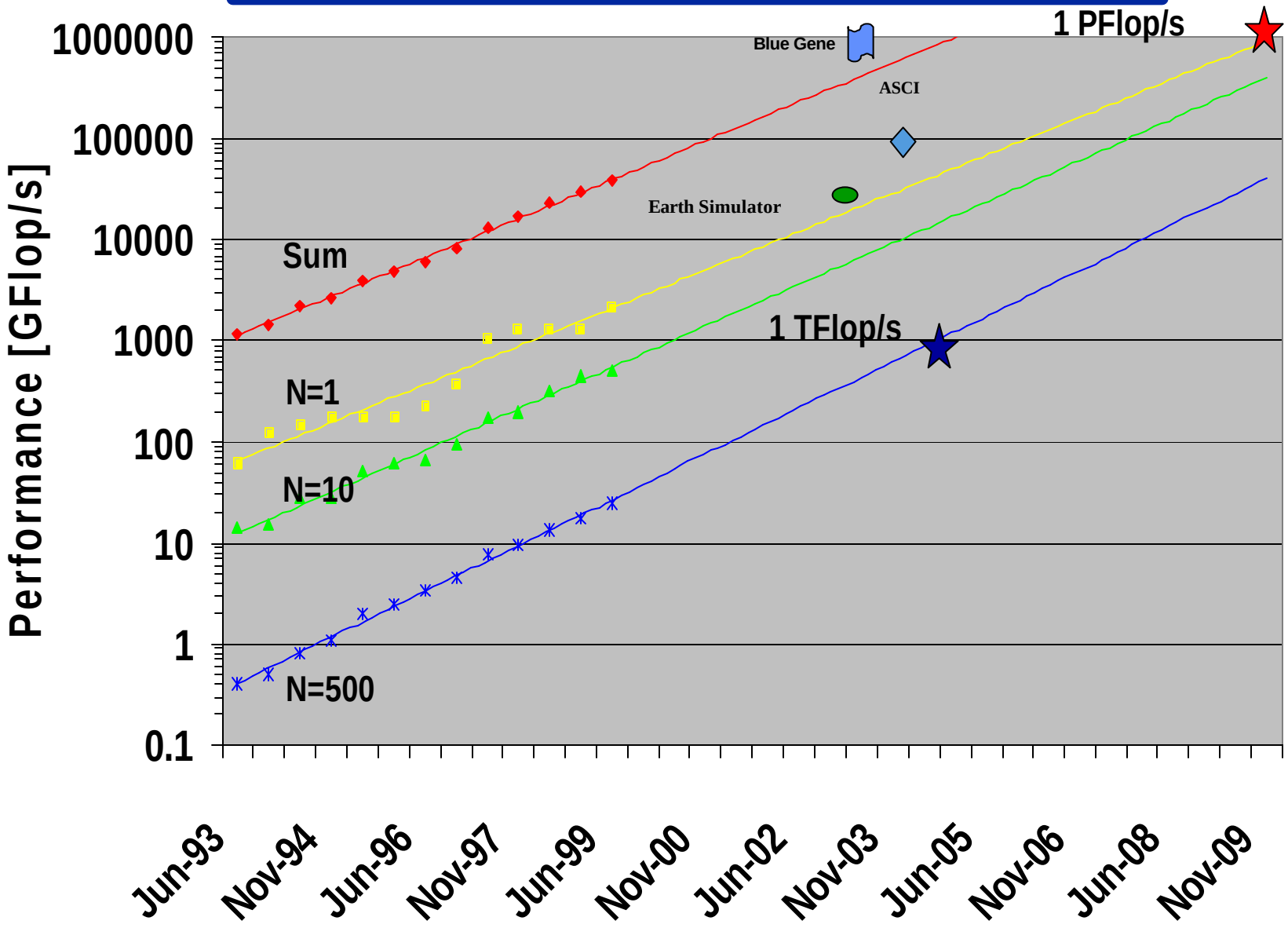
Supernova simulation:

- Critical need to better understand Type 1a supernovas, since these are used as “standard candles” in calculating distances to remote galaxies.
- Current models are only 2-D.
- Initial 3-D model calculations will require 2,000,000 CPU-hours per year, on jobs exceeding 256 Gbyte memory.
- Future calculations 10 to 100 times as expensive.

Analysis of cosmic microwave background data:

MAXIMA data	5.3×10^{16} flops	100 Gbyte mem
BOOMERANG data	1.0×10^{19} flops	3.2 Tbyte mem
Future MAP data	1.0×10^{20} flops	16 Tbyte mem
Future PLANCK data	1.0×10^{23} flops	1.6 Pbyte mem

Top500 Trends



Top500 Data Projections

- First 100 Tflop/s system by 2005.
- No system under 1 TFlop/s will make the Top500 list by 2005.
- First commercial Pflop/s system will be available in 2010.

For info on Top500 list, see http://www.top500.org

The Japanese Earth Simulator System

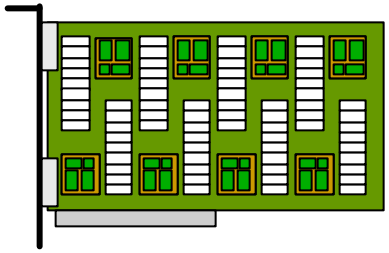
System design:

- Architecture: Crossbar-connected multi-proc vector system.
- Performance: 640 nodes x 8 proc per node x 8 Gflop/s per proc = 40.96 Tflop/s peak
- Memory: 640 nodes x 16 Gbyte per node = 10.24 Tbyte.

Sustained performance:

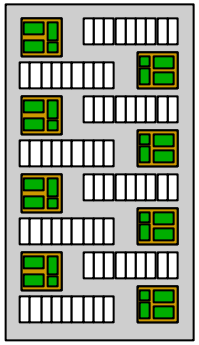
- Global atmospheric simulation: 26.6 Tflop/s.
- Fusion simulation (all HPF code): 12.5 Tflop/s.
- Turbulence simulation (global FFTs): 12.4 Tflop/s.

IBM's Blue Gene/L Project



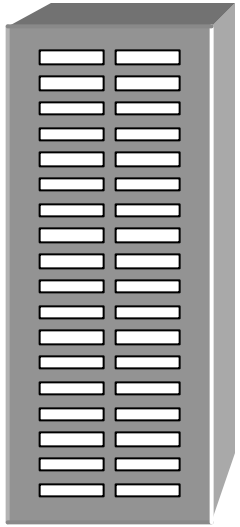
22GF on a Card

8 nodes
(2x2x2)
22.2/44.8 GF/s



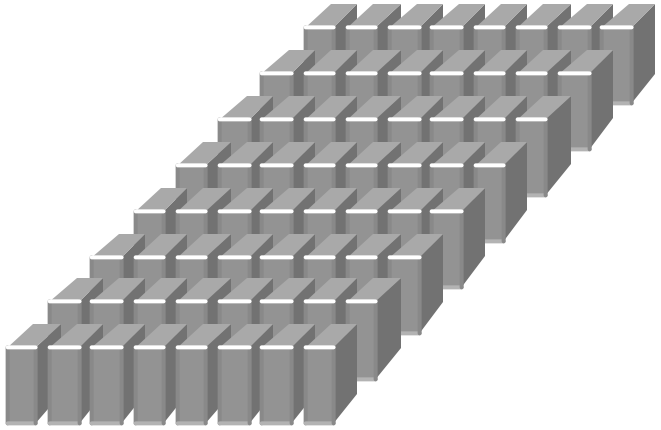
BOARD

8 nodes
(2x2x2)
22.2/44.8 GF/s
2.08 GiB*



CABINET

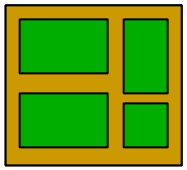
128 boards
(8x8x16)
2.9/5.7 TF/s
266 GiB*
15 kW



SYSTEM

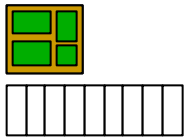
64 cabinets
(32x32x64)
180/360 TF/s
16 TiB*
~1 MW
2500 sq.ft.

~12mm



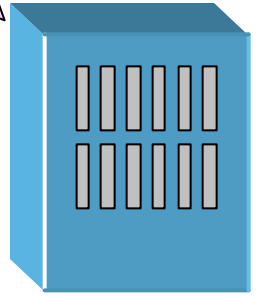
COMPUTE ASIC

2 processors
2.8/5.6 GF/s
4 MiB*



NODE

2 processors
2.8/5.6 GF/s
256 MiB*
15 W



1TF in a Box

64 boards
(8x8x8)
1.4/2.9 TF/s

Other Future High-End Designs

- Processor in memory
  - Currently being pursued by a team headed by Prof. Thomas Sterling of Cal Tech.
  - Seeks to design a high-end scientific system based on special processors with embedded memory.
  - Advantage: significantly greater processor-memory bandwidth.
- Streaming supercomputer
  - Currently being pursued by a team headed by Prof. William Dally of Stanford.
  - Seeks to adapt streaming processing technology, now used in game market, to scientific computing.
  - Projects 200 Tflop/s, 200 Tbyte system will cost $10M in 2007.

Petaflops Computing

- 1 Pflop/s (10^{15} flop/sec) in computing power.
- Between 10,000 and 100,000 individual CPUs.
- Between 10 Tbyte and 1 Pbyte main memory
  (= 100x the UC Berkeley library).
- Between 10 and 100 Pbyte on-line mass storage.
- If built today, a petaflops system would cost $1 billion and consume 100 Mwatts of electric power.
- Programming challenge: 10^8-way concurrency at all significant steps of a computation.

Future Applications for Petaflops Systems

- Weather forecasting.
- Business data mining.
- DNA sequence analysis.
- Protein folding simulations.
- Inter-species DNA analyses.
- Medical imaging and analysis.
- Nuclear weapons stewardship.
- Multiuser immersive virtual reality.
- National-scale economic modeling.
- Climate and environmental modeling.
- Molecular nanotechnology design tools.
- Cryptography and digital signal processing.

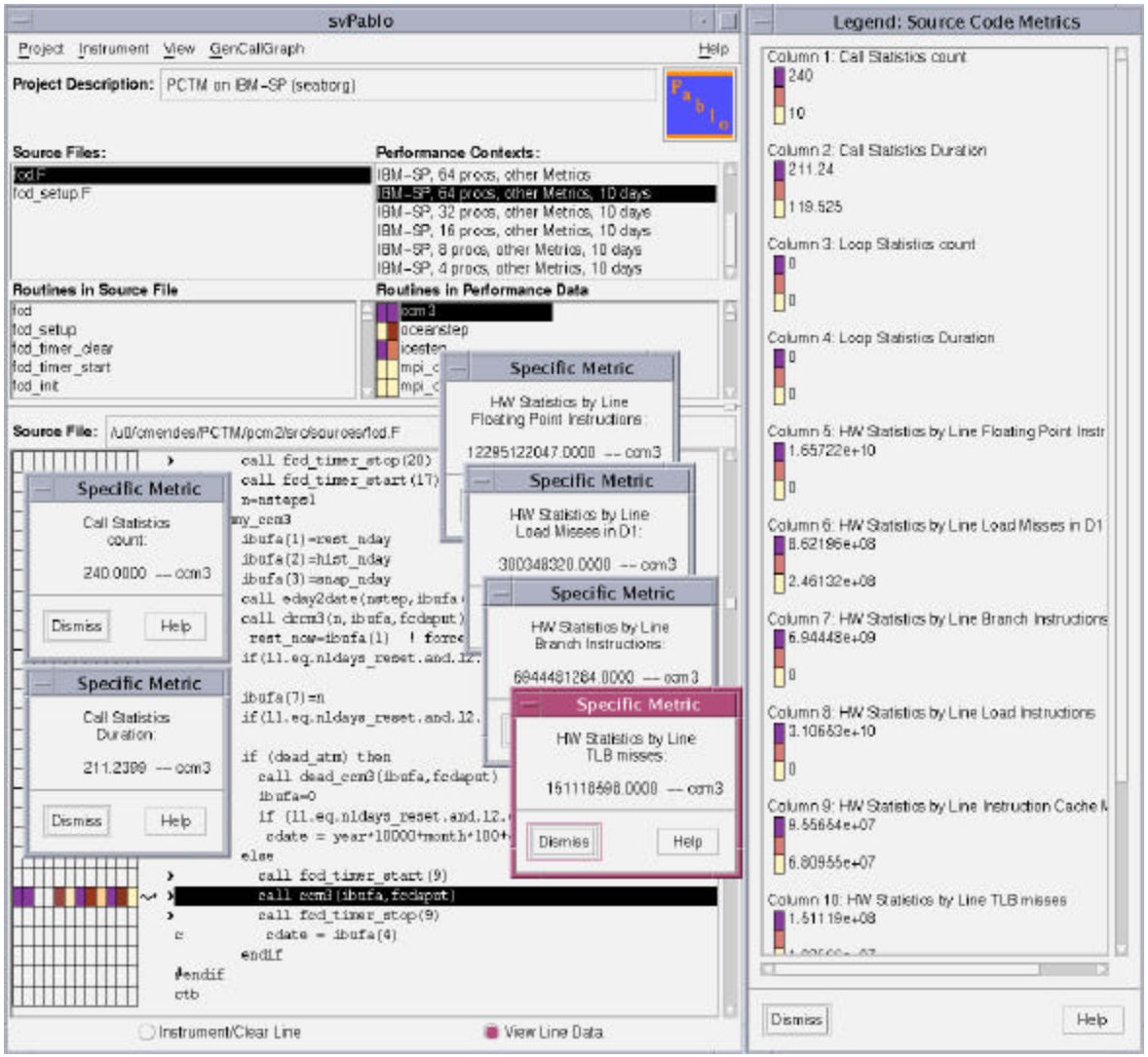
Questions for Future High-End Computing

- Can systems with 10,000 to 100,000 or more processors deliver acceptably scaled performance?
- How can procurement teams intelligently select systems that are an order of magnitude larger than any system currently fielded?
- Will fundamental algorithm concurrency be a limiting issue at this scale?
- Will floating-point arithmetic accuracy be an issue (i.e., when will 128-bit arithmetic be required)?
- Can existing system software be used at this scale?
- Will new programming models be required?

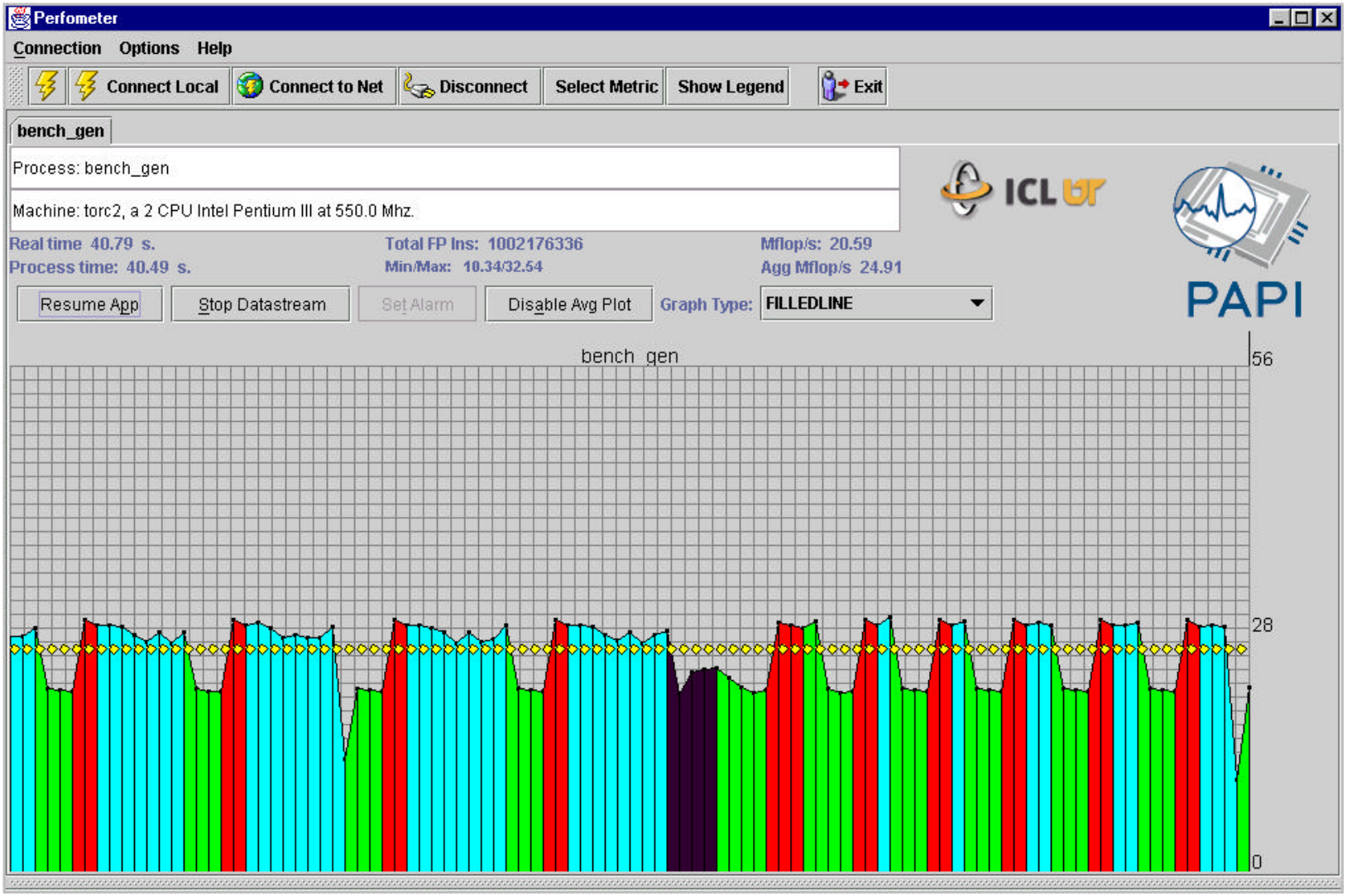
The Performance Evaluation Research Center (PERC)

- One of five Integrated Software Infrastructure Centers funded through the DoE SciDAC program.
- Research thrusts:
  - Development of improved tools for performance monitoring and code tuning.
  - Studying the performance characteristics of specific large-scale scientific codes.
  - Development of tools and techniques for performance modeling.
  - Development of semi-automatic facilities for improving performance.

User Tools: SvPablo

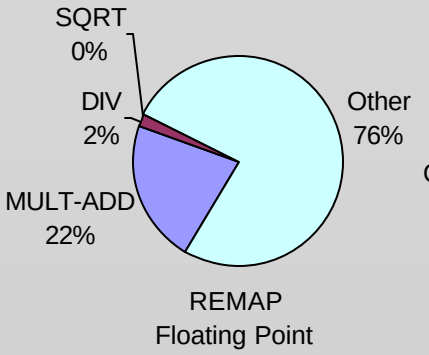


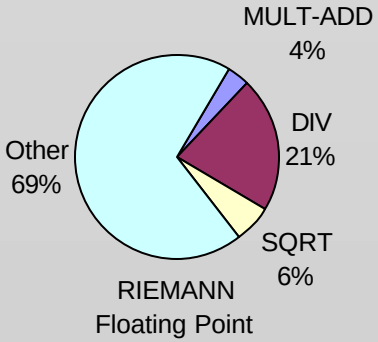
PAPI Perfometer Interface

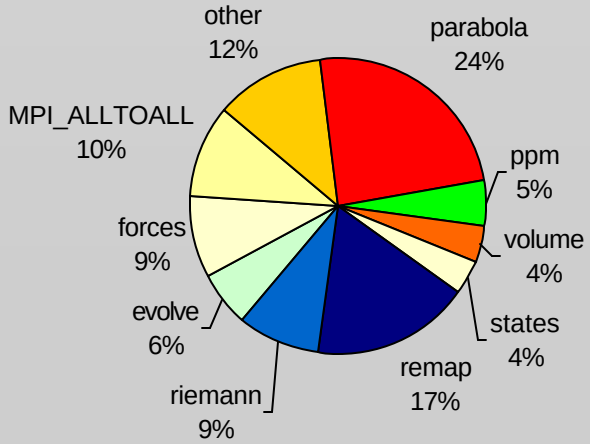


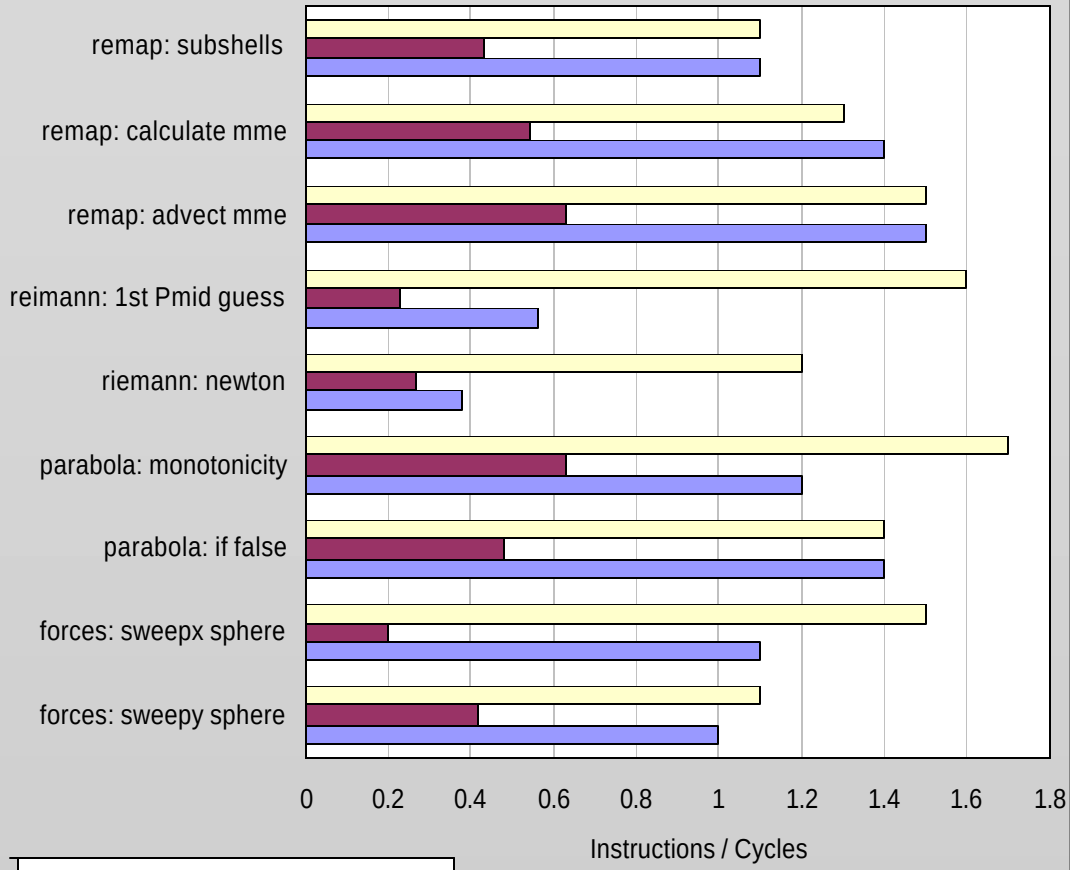
Performance Analysis
EVH1 (high-energy physics)

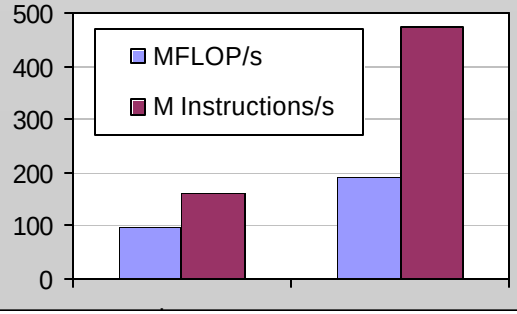
Aggregate performance measures over all tasks for a .1 simulation-second run. Collected with PAPI on an IBM SP (Nighthawk II / 375MHz).

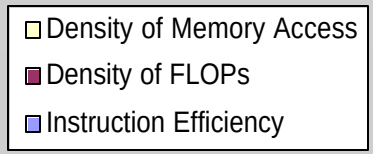












PERC Performance Modeling

- *Application signature* tools characterize applications independent of the machine where they execute.
- *Machine signature* tools characterize computer systems, independent of the applications.
- *Convolution* tools combine application and machine signatures to provide accurate performance models.
- *Statistical models* find approximate performance models based on easily measured performance data.
- *Phase models* analyze separate sections of an application, enabling overall performance predictions.
- *Performance bound* tools determine ultimate potential of an application on a given system.

Sample Modeling Results

# CPUs	Real Time	Predicted Time	% Error
2	31.78	31.82	0.13
4	29.07	31.27	7.57
8	36.13	33.72	6.67
64	44.91	43.91	2.23
96	48.87	47.15	3.52
128	52.88	52.46	0.79

Key Challenges in the Performance Research Field

- Scaling performance monitoring tools to many thousands of processors.
- Handling the exploding volume of trace data.
- Visualizing performance results.
- Developing performance modeling tools accurate enough to predict performance in the 10,000-100,000 processor arena.
- Extension of tools and modeling techniques to cover emerging architectures (Cray X1, ESS, PIM, BG/L).